Tigrinya Part-of-Speech Tagging with Morphological Patterns and the New Nagaoka Tigrinya Corpus

Yemane Keleta Tedla
Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka
Niigata, 940-2188 Japan

Kazuhide Yamamoto
Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka
Niigata, 940-2188 Japan

Ashuboda Marasinghe
Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka
Niigata, 940-2188 Japan

ABSTRACT
This paper presents the first part-of-speech (POS) tagging research for Tigrinya (Semitic language) from the newly constructed Nagaoka Tigrinya Corpus. The raw text was extracted from a newspaper published in Eritrea in the Tigrinya language. This initial corpus was cleaned and formatted in plaintext and the Text Encoding Initiative (TEI) XML format. A tagset of 73 tags was designed, and the corpus for POS was manually annotated. This tagset encompasses three levels of grammatical information, which are the main POS categories, subcategories, and POS clitics. The POS tagged corpus contains 72,080 tokens. Tigrinya has a unique pattern of root-template morphology that can be utilized to infer POS categories. Subsequently, a supervised learning approach based on conditional random fields (CRFs) and support vector machines (SVMs) was applied, trained over contextual features of words and POS tags, morphological patterns, and affixes. A rigorous parameter optimization was performed and different combinations of features, data size, and tagsets were experimented upon to boost the overall accuracy, and particularly the prediction of POS for unknown words. For a reduced tagset of 20 tags, an overall accuracy of 90.89% was obtained on a stratified 10-fold cross validation. Enriching contextual features with morphological and affix features improved performance up to 41.01 percentage point, which is significant.

General Terms
natural language processing, part-of-speech tagging

Keywords
Semitic languages, Tigrinya corpus, Tigrinya part-of-speech tagging, morphological patterns

1. INTRODUCTION
Parts-of-speech are lexical categories of words such as verbs, nouns, and adjectives. The process of labeling a word with its appropriate lexical category is known as POS tagging. A word's POS role may differ depending on its lexical information or the syntactic arrangement of the surrounding context. POS tagging is a fundamental process in the stages of natural language processing (NLP). As noted in [24], although resource rich languages such as English have well-developed language tools, low-resource languages suffer from low grade or absence of electronic data support to pursue NLP research. The Tigrinya language is one of the low-resource languages. Tigrinya belongs to the Semitic language family of the Afro-Asiatic phylum, along with Hebrew, Amharic, Maltese, Tigre, and Arabic. Tigrinya claims an estimated 7 million speakers in Eritrea and northern Ethiopia. Unlike major Semitic languages, which enjoyed relatively widespread NLP research and resources, the Tigrinya language was largely ignored in NLP-related research due to the absence of a Tigrinya corpus. Although a few electronic dictionaries and a lexicon of automatically extracted Tigrinya words exist, a text corpus that has linguistic information is not available for the public. Given this circumstance, new research needed to initiate the NLP research from the foundation by constructing a corpus, and thereby language tools, for the advancement of information access in Tigrinya. Specifically, the objectives of this research are constructing a new POS-tagged corpus and developing a POS tagger using two state-of-the-art supervised machine learning algorithms, augmented with morphological features of the Tigrinya language.

The remaining content is organized into seven sections. Following the brief introduction of the research, section two discusses the linguistic characteristics of Tigrinya, with a focus on POS ambiguity challenges. A survey of earlier POS-tagging research among Semitic languages is presented in section three. Section four briefly introduces the new POS-tagged corpus, named the Nagaoka Tigrinya Corpus (NTC), and section five continues to outline the process of extracting informative morphological patterns. Subsequently, a new experimental setup and results are presented in section six. Finally, conclusion and future direction are given in section seven.

2. TIGRINYA LANGUAGE
2.1 A "word" in Tigrinya
Semitic languages, in general, are characterized by rich inflectional and derivational morphology, which generates numerous variations of word forms. The presence of a variable word form or morphology is quite different from other language families, such as Indo-European, Nilo-Saharan, or others. The distinguishing feature of Semitic languages lies in the 'root-template' morphological pattern that is often composed of triliteral roots. The verb roots in Semitic languages comprise a sequence of consonants, whereas the templates are the patterns of vowels that are intercalated in between these consonants, forming various stems. Accordingly, looking at the inflection of verbs it is possible to extract linguistic information such as tense-aspect-mood, number, gender, person, object suffix, negation and case.

For example, the word ፈለጠ /feliTu/ meaning '[he] knew'[1], is the intercalation of the root 'flT' and the gerundive template 'eiu'. Generally, words related to the concept 'know' would be generated by intercalating 'flT' and other vowel templates

[1] The pronoun 'he' is not explicitly mentioned, but it is indicated in the inflection of feliTu. We use the notation '[pronoun]' for such cases.

with possible affixation of prefixes and suffixes. A few examples of 'flT' productions are given in table 2.

Although Tigrinya words are delimited by spaces, a word or a token may be a sequence of grammatical morphemes that coexist affixed to each other. For example, the single token 'እንተዘይሓተትካዮ' /IntezeyIHatetIkayo/[2] can be translated to 'if you did not ask him', expressed in six English words. Figure 1 depicts the analysis of morphological and POS information that are embedded in this token. Accordingly, we may extract from the token, the POS category (VERB), subcategory (perfective), the inseparable conjunction proclitics (CON), negation prefix (REL+NEG), subject pronoun (SUBJ.PRO), and object pronoun (OBJ.PRO) suffixes.

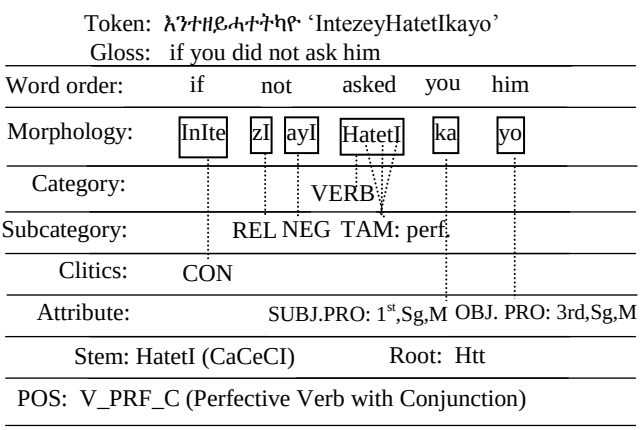

Figure 1: Morphology and POS in Tigrinya

In a pioneering work for the analysis and generation of Tigrinya morphology, Gasser describes the arrangement of grammatical morphemes of Tigrinya verbs in the following order [9].

(prep|conj)(rel)(neg)sbjSTEMsbj(obj)(neg)(conj)

The slots preceding the 'STEM' are prefixes that indicate a preposition (prep), a conjunction (conj), and a relativizer (rel). The negation 'neg' morpheme (ኣይ...ን /ayI...nI/) circumfixes the STEM. For example, Hatete, '[he] asked,' when negated becomes ayI-Hatete-nI, '[he] did not ask' (without a hyphen). Like the other Semitic languages, Tigrinya is a highly inflected language. Tigrinya verbs are inflected for gender, person, number, case, tense-aspect-mood (TAM), and so on. These inflections, which involve complex affixation of 'sbjSTEMsbj' rules, also enforce subject-verb agreement. For example, considering the perfective STEM /HatetI/, roughly 'asked,' HatetI-ku means 'I asked,' HatetI-na means 'we asked,' and so on. Therefore, it is possible to recover the pronoun by looking at the verb inflection because verbs must agree with subjects. Similarly, the slots following the STEM are suffixes that indicate object pronoun (obj) and conjunction elements. Further discussion of morphological features and their impact in boosting performance of the POS taggers is available in section six. Additional characteristics of the Tigrinya language are described in the following subsections, with emphasis on related POS ambiguity.

[2] The transliteration mapping in this paper uses the SERA scheme (*http://www.yacob.org*). However, the upper case 'I' was used to exclusively mark the epenthetic vowel, traditionally known as 'sadIsI'.

2.2 Writing System

The Ge'ez language was the proto-family for the Tigrinya, Amharic, and Tigre languages, and currently these languages share the Ge'ez alphabets. These three languages replaced Ge'ez as vernacular languages, while Ge'ez remained confined to liturgical and archival domains. Hence, Tigrinya is one of the few African languages with an indigenous writing system.

Unlike Arabic and Hebrew scripts Ge'ez script is written from left to write. Ge'ez script is an *abugida* system in which each letter (alphabet) represents a consonant-vowel (CV) syllable. Accordingly, Tigrinya identifies seven vowels usually called 'orders' [17]. There are also a few alphabets which are variants of some of the main 35 alphabets but have only five orders. Altogether, about 275 symbols make up the Tigrinya alphabet chart known as 'Fidel'.

In connection with POS tagging, the phenomenon of gemination (lengthening of consonants) creates ambiguity in the POS of words. This ambiguity results from the absence of notation symbols in Ge'ez alphabets, such as orthography of the European languages to represent gemination of consonants. For example, the word ሰበረ /sebere/ can mean '[he] broke' or it can be a type of a legume if '**be**' in 's**be**re' is geminated. Furthermore, the widespread use of cliticized words may pose serious problems in POS tagging because of the orthographic variation it creates. During Tigrinya cliticization, certain unpronounced characters of a word are omitted and replaced by apostrophes. For example, the compound word መምህር'ዩ /memIhIrI'yu/, '[he] is a teacher,' is a combination of the noun 'memIhIrI' and the auxiliary verb 'Iyu.' However, during fussion of these words, the first character of the auxiliary verb is omitted by cliticization. The proposed system includes some character recovery rules and has normalized the corpus in order to reduce these types of orthographic variations. Orthographic variations may aggravate the out-of-vocabulary problem that often occurs with low-resource languages. However, the character recovery process is not always straightforward because there are some combinations that require more contextual or semantic knowledge to determine the proper missing character. For example, the word ክመይ'ላ /kemeyI'la/ could be resolved to either /kemeyI ala/, 'how is [she],' or /kemeyI ila/ 'how did [she].' Issues with recovery of such ambiguous cases will be resolved in forthcoming enhancements of the corpus quality.

2.3 POS Ambiguity in Tigrinya

POS ambiguity may arise from the behavior of some words to assume various POS roles according to their lexical information and the surrounding context. Some cases of ambiguity in Tigrinya are discussed in this section.

Tigrinya demonstrative pronouns and demonstrative adjectives may be ambiguous, depending on the word being modified [17]. One could consider the statements (1) እዚ ናተይ እዩ /Izi natey Iyu/, 'this is mine,' and (2) እዚ ቤት እዚ ናተይ እዩ /Izi bEtI Izi natey eyu/, 'this house is mine.' In (1), Izi/this' functions as a pronoun, whereas in (2), 'this' is modifying the noun 'house' and is hence an adjective. In Tigrinya, demonstrative adjectives tend to repeat themselves in the pattern 'Izi NOUN Izi.' Furthermore, an adjective may take the role of a noun, a relative verb, a pronoun, a proper noun, or an interjection in a sentence [22]. Inflection and affixation of Tigrinya words may also render lexically ambiguous words. The prefix ብ /bI/ 'by' for instance, creates ambiguity because in certain cases it turns the role of a noun into an adverb [11]. In the word ብመምህር /bImemIhIrI/ 'by a teacher',

the prefix 'bI' is a preposition. However, for the word ብትብዓት /bItIbIAtI/ 'bravely', the word assumes the role of an adverb, specifically in the event the word modifies a verb. However 'bI' is not always a prefix or adverb creator but sometimes part of a word too. ብስራት /bIsIratI/ 'good news' which can appear both as a noun and proper noun has 'bI' as a part of the word not a morpheme. This ambiguity can be partly solved by a combination of stemming and lexicon look-up to check if the stem exists in the lexicon or alternatively looking at the context for disambiguation information such as the part-of-speech of the word being modified. In our research, we applied the latter alternative to disambiguate the POS based on its context.

One of the features of Tigrinya as a Semitic language is that it has two 'tenses' (perfective and imperfective) but several 'aspects' (causative, reflexive, reciprocal, and so on) and the imperative/jussive 'mood' [17]. Other such tenses are 'future tense,' which is expressed by combining one of these tenses with auxiliary verbs. This distinction becomes useful when using syntactic information to understand the arrangement of words for POS disambiguation. For example, the phrase ዘሊሉ ሃደመ /zelilu hademe/, '[he] escaped [by] jumping,' describes the way the 'escape' action was performed. Therefore, a gerundive verb may function as an adverb when followed by a perfective verb [11].

The declension of nouns in Tigrinya happens for gender, number, case and definiteness. However noun declension does not follow a regular pattern in almost 75% of the time [14]. Similarily declension of adjectives takes place for gender and number. In general, the complexity of Tigrinya grammar leads to several ambiguities related to all parts of speech.

3. RELATED WORKS

This section reviews earlier POS-tagging works conducted in Semitic languages, which are categorized under the same language branch as the Tigrinya language. There are two approaches in POS tagging for Semitic languages. The first approach focuses on the discrete nature of words without any morpheme-level segmentation, while the second follows a decomposition strategy and works on segments or morphemes of words. Therefore, two trajectories prevail in the tagging of Semitic languages that lead to either words with complex POS tags or a sequence of segments with simple POS tags. The latter requires choosing among multiple ambiguous segments [2]. [13] pioneered the first POS tagging for the Arabic language with a corpus containing 50,000 tagged words collected from a newspaper. Following [13], separate tagging for the Arabic language emerged over the decades, and currently, Arabic corpora with millions of words are available. Arabic Treebank, containing 1 million words [5], and the Arabic Newswire part-1 [15], comprising 76 million tokens and over 666,094 unique words, were compiled at the University of Pennsylvania. Methodologically, the majority of recent studies have applied segmentation-based techniques for Arabic POS tagging [7, 16, 20]. [7] reported an SVM-based POS tagger that performs morphological segmentation of words followed by POS tagging. Interestingly, a comparative study of segmentation-based and segmentation-free methods by [19] reports better results without segmentation (94.74% vs. 93.47%). In contrast, [2] examined the problem of word tokenization for Hebrew POS tagging and argued that segment-level tagging is better suited for POS tagging of Semitic languages in general, as it suffers less from data sparseness. A recent work investigated modeling POS tagging as an optimization problem using the genetic algorithm [1]. POS-tagging research for Amharic languages was mostly driven by a 210K-word news corpus that was tagged with 30 POS tags [6] Using this corpus, [8] reported various experiments applying Trigram'n'Tags (TnT), SVM, and maximum entropy algorithms. Furthermore, [10] identified inflectional and derivational patterns of Amharic words as features and improved tagging accuracy up to 90.95% using CRFs. The research on Tigrinya NLP has not advanced much. However, a notable achievement on morphological analysis and generation of Tigrinya, Amharic, and Oromo has been reported by [18] This research applied finite state transducers with feature selection to handle template morphology and long-distance dependencies in Tigrinya verb morphotactics. Moreover, a recent work by [21] reported a hybrid stemmer with a performance of 89.3%. Some electronic dictionaries[3], input method editors[4], and a large Tigrinya lexicon from the Crúbadán project[5] are also available. A very recent project is working on web-crawling a number of languages in the horn of Africa including Tigrinya. The corpora created can be accessed online through concordancing[6].

4. THE NAGAOKA TIGRINYA CORPUS

4.1 Data Collection

The raw text of the corpus was collected from the issues of a national newspaper called *Haddas Ertra*, published in Eritrea. Each issue is uploaded to the official website of the Ministry of Information at *www.shabait.com* in PDF format. *Haddas Ertra* was chosen because, comparatively, it is the most diversified source of Tigrinya text in terms of topic, domain, and stylometery. The text content includes, among others types, news, editorial, reportage, commentary, interviews, stories, and biographies.

The initial repository of articles was created by automatically downloading the *Haddas Ertra* issues published at *shabait.com*. Next, raw text was extracted and preprocessed with rule-based tools for cleaning and normalizing the text. The size of the text corpus after the preprocessing phase is around 9.1 million tokens. A small portion of the clean corpus was manually selected at random for manual POS tagging. Articles include topics from health, education, agriculture, business, sports, health, social issues, history, culture and literature. The text that was manually annotated contains 72,080 tokens. The presented research employs this tagged corpus for the development of a POS tagger in Tigrinya.

4.2 Corpus Design

The design and development of NTC, including data collection, preprocessing, tagset design, and annotation, is described at [27]. The current POS study is an extension of the corpus construction research. Table 1 lists all the labels of the POS tags and the overall distribution of the 12 major tags is depicted in figure 2.

[3] *http://www.hdrimedia.com/, http://www.memhr.org/dic/, http://www.geezexperience.com/dictionary/*
[4] *http://keyman.com/tigrigna/, http://geezlab.co/*
[5] *http://crubadan.org/languages/ti*
[6] *http://habit-project.eu/wiki/CorporaAndCorpusBuilding*

Table 1: The full tagset: labels with _C, _P, _PC ending idicates clitics of conjunction, preposition, or both respectively. Example, N_C refers to a noun with conjunction, and so on .

Category	Label
Noun	N,N_C, N_P,N_PC
Verbal	N_V,N_V_C,N_V_P,N_V_PC
Proper	N_PRP,N_PRP_C,N_PRP_P,N_PRP_PC
Pronoun	PRO,PRO_C,PRO_P,PRO_PC
Verb	V,V_C,V_P,V_PC
Perfective	V_PRF,V_PRF_C,V_PRF_P,V_PRF_PC
Imperfective	V_IMF,V_IMF_C,V_IMF_P,V_IMF_PC
Imperative	V_IMV,V_IMV_C,V_IMV_P,V_IMV_PC
Gerundive	V_GER,V_GER_C,V_GER_P,V_GER_PC
Relative	V_REL,V_REL_C,V_REL_P,V_REL_PC
Auxiliary	V_AUX
Adjective	ADJ,ADJ_C,ADJ_P,ADJ_PC
Adverb	ADV,ADV_C,ADV_P,ADV_PC
Preposition	PRE,PRE_C,PRE_P,PRE_PC
Conjunction	CON,CON_C,CON_P,CON_PC
Interjection	INT,INT_C,INT_P,INT_PC
Numeral	NUM
Cardinal	NUM_CD,NUM_CD_C,NUM_CD_P, NUM_CD_PC
Ordinal	NUM_OR,NUM_OR_C,NUM_OR_P, NUM_OR_PC
Punctuation	PUN
Foreign Word	FW
Unclassified	UNC

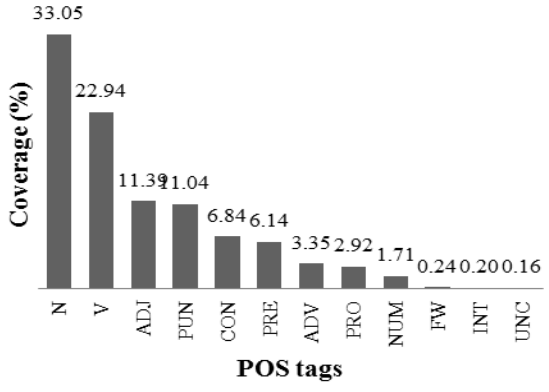

Figure 2: Distribution of tags and corpus statistics

The statistical summary of the NTC corpus is as follows:

Data source: Haddas Ertra newspaper
Articles: 100, from around 10 Topics
Corpus size: 72,080 tokens
Sentences: 4656 (avg. 15 tokens/sent)
Unique words: 18,740 (26%)
All Tags: 73
Token-type ratio: 3.85
Hapaxes: 12,510 (17%)

Many Tigrinya linguists classify Tigrinya POS into eight major categories [14, 17, 22]. These are verbs, nouns, pronouns, adjectives, adverbs, conjunctions, prepositions and interjections. Other Tigrinya POS tags proposed by [25] reduced POS types into five categories namely; verbs, nouns, adjectives, prepositions and adverbs. Earlier, [14] introduced definite and indefinite articles as separate POS Tigrinya. The proposed tagset follows the first classification which is the most agreed and has better diversity of tags. A total of 73 tags or 'tagset-1' that are divided into three levels of annotation were recognized in due process. Level-1 refers to the main 12 categories, level-2 details some selected subcategories and level-3 annotates preposition and conjunction clitics (table 1). The tagset is similar to a previous work by [6] with enhancements primarily to include subcategories of verbs. NTC was originally tagged with tagset-1 and later reduced to 'tagset-2', preserving only 20 tags to reduce data sparseness. The experiments and results section discusses the impact of both tagsets on the performance of the presented taggers. NTC has been released for the public[7] in the hope of advancing NLP research for Tigrinya.

5. EXTRACTING PATTERNS

In the English language, 98% of the cases for the suffix 'able' were found to be adjectives in the *Wall Street Journal* part of Penn Treebank [3]. This illustrates the possibility of exploiting suffix rules for predicting POS. Morphological features may be extracted statistically or hand-picked by experts. [21] showed that a few linguistically motivated suffixes can result in improved generalization in English. On the other hand, [26] implemented a maximum entropy Markov model to automatically extract morphological features. This improved the accuracy of tagging unknown words from 61% to 80%, tested on Chinese Treebank 5.0. "Unknown words" are the words that were not seen during the training but were present in the test data. As mentioned previously, Tigrinya words possess rich morphological information embedded in the form of prefixes, infixes, and suffixes. Figure 1 illustrates the morphological and POS information that can be encoded in the words of Tigrinya. Tigrinya verbs use suffixes to mark the conjugation of personal pronouns. For example, the perfective verb pattern CeCeCe is conjugated by inflecting the stem CeCeC with one the following personal pronoun suffixes (C = represents any Tigrinya consonant).

CeCeC+(e | etI | a | u | ka | ki | kumI | kInI | na| ku)

For instance, the verb /sebere/ can be decomposed into seber+ e, which translates to 'he broke.' Likewise, the pattern of the stems for the remaining verbs in the /sebere/ family are all distinct. Specifically, the pattern of the imperfective stem is -CeCC- (-sebr-), the imperative stem pattern is -CCeC (-sber), and the gerundive stem has a CeCiC- (sebir-) pattern. For further illustration, some of the most informative patterns based on a frequency distribution of gerundive verbs (V_GER) are listed in table 2. The V_GER proportion in the second column gives the percentage of a particular V_GER pattern compared to a total of 533 extracted V_GER patterns (including inflected patterns). The pattern 'CeCiCu' is the most dominant gerundive stem in the corpus (31.1%), with the

[7] The Nagaoka Tigrinya Corpus is publicly available at *http://eng.jnlp.org/yemane/ntigcorpus.*

suffix '-u' indicating masculine, singular, and third-person attributes. In addition to the pronoun suffix that is present in all verbs, stems with prefixes such as 'te-' and 'a-' are also frequently found in the corpus. The statistics reveal that 17.8% of the gerundive verb inflections are prefixed with 'te-,' which forms passive voice stems from Tigrinya perfective and imperfective verbs. Similarly, the pattern of the causative stem 'aCICeCe,' which is prefixed by the morpheme 'a-,' is retrieved as part of the most informative patterns of gerundive verbs. These types of morphological clues are very important in correctly predicting the POS of Tigrinya words.

Table 2: Some patterns of gerundive verbs (V_GER)

V_GER patterns	V_GER (%)	Examples
CeCiCu	31.1	feliTu 'he knew'
CeCeCiCu	17.8	tefeliTu 'it was known'
CeCaCiCu	15.9	tefaliTu 'to know each other'
CeCiCICa	14.6	feliTIka 'you knew'
aCICiCu	11.8	afIliTu 'made something known'
CeCiCoCI	11.3	feliTomI 'they knew'
CeCiCa	10.5	feliTa 'she knew'

Morphological patterns of Semitic languages have proven to be strong POS indicators of words. A recent research on Amharic POS tagging investigated morphological and derivational rules of Amharic and employed consonant and vowel patterns as features. The approach improved the accuracy up to 90.95% which is better than preceding POS research on Amharic.In the proposed feature set for Tigrinya, the morphological patterns were automatically extracted, and the contextual features were enriched to include succeeding words in the feature window. As reported in table 4, the proposed feature sets improved the accuracy of unknown words significantly.

6. EXPERIMENTS AND RESULTS

6.1 Experimental Setup

6.1.1 Data

The NTC corpus was used for training and testing. There are two versions of the corpus available based on tagset-1 (the full tagset of 73 tags) and tagset-2 (the reduced tagset, containing 20 tags). On average, almost 81% of the training data are known words and 18% are unknown words. Experiments included various sizes of the corpus to analyze its effect on accuracy (table 5).

6.1.2 Features

The full set of the proposed features includes a rich set of contextual and lexical features. The contextual features span a window of two words and two POS tags preceding the focus word, the focus word, and two words succeeding the focus word. Additionally, lexical features are extracted from the focus word's affixes, which comprise prefixes of one to six characters in length, consonant-vowel patterns of the word (infixes), and suffixes of one to five characters in length. The length of characters is decided from the best results of a grid search of character combinations.

6.1.3 Algorithms

One of the two important objectives of the research was initiating construction of a POS-tagged Tigrinya corpus; therefore, supervised learning was employed. Two different learning approaches were used in the experiments. The first is conditional random fields (CRFs), a sequence-labeling algorithm that takes contextual dependence into account. CRFs are preferred over other models because they offer improved sequence labeling solutions. Firstly, the strict Markovian assumption in hidden Markov models (HMMs) is relaxed in CRFs. CRFs also solve the bias label problem that exists in maximum entropy Markov models (MEMMs) by training to predict the whole sequence correct instead of training to predict each label independently [12]. In the second approach, POS tagging is modeled as a classification problem and uses an SVM classifier [4] in a one-versus-the-rest multiclass scheme[8]. Both approaches are among the methods that have been successfully applied to develop state-of-the-art POS taggers[9].

6.1.4 Hyperparameter Optimization

A grid search was applied to find the optimum value of the *C* parameter, *l1* and *l2* regularizations, and the kernel function for the SVM-based tagger. Accordingly, the best estimation was when *C*=1, using *l2* regularization and the linear kernel. Similarly, a randomized grid search optimization for the CRF-based experiments based on the *'lbfgs'* kernel was applied. The best estimation was found when *C1*=0.5978 and *C2*=0.1598.

6.1.5 Evaluation

The overall accuracy is calculated from test data by computing the percentage of correctly predicted tags to the true annotations.

$$Accuracy = \frac{Number\ of\ Correctly\ tagged\ tokens}{Total\ number\ of\ tokens}$$

A stratified 10-fold cross validation (CV) scheme evaluated the performance of the estimators. CV setup is particularly useful in a low-resource scenario when the data for training and testing is not sufficient. The stratified CV creates a balanced dataset by distributing approximately the right proportion of tags into each of the 10 folds. Therefore, each fold may be regarded as representative of the whole data. In addition, standard metrics of precision, recall, and F1-score were utilized to assess the system's performance of tag assignments for individual POS tags. Finally, error analysis analyzes the performance strengths and weaknesses of the taggers on unknown words and specific tags.

6.1.6 Baseline

For the baseline, only the current word was considered as the feature, regardless of its context, and unknown words were tagged as nouns. On tagset-1, the SVM tagger yielded a baseline accuracy of 74.05%, while the CRF tagger performed slightly better at 76.15%. All the other results obtained outperform these baselines (table 3).

[8] Experiments, optimizations, and evaluations were carried out using scikit-learn machine learning tools. For CRF, the sklearn-crfsuite package was used, which is available at http://sklearn-crfsuite.readthedocs.io/en/latest/.

[9] *http://www.aclweb.org/aclwiki/*

6.2 Overall Results

In general, the CRF tagger slightly outperforms the SVM-based tagger. Table 3 shows the overall performance of the taggers by tagset and affix features over the data size from both tagsets. Accordingly, the highest CRF score is 90.89% and that of SVM is 89.92% on tagset-2. The difference of 0.01% in percentage point may seem trivial but a *p*-value significance test proves that the difference is quite significant ($p = 0.002 < 0.05$). POS tagging is a sequence-labeling task, therfore, CRF has the advantage of learning relations from surrounding context. This may be the reason why CRF-based tagger tends to outperform the SVM-based tagger in most of the experiments. The pattern features are unique to Tigrinya (and other Semitic languages). The disambiguation of the four types of verbs (perfective, imperfective, imperative, and gerundive), nouns, and adjectives is largely backed up by these infix features. According to table 3, in both experiments of CRF and SVM, patterns features boost performance by more than 5.5 percentage point from the baseline.

6.3 The Effect of Tagset Design

The rich inflectional and derivational morphology of Tigrinya generates a large number of words with various grammatical information, such as gender, number, case, and so on. In addition, there are also clitics of prepositions and conjunctions that are affixed to words in Tigrinya. As mentioned in earlier sections, the initial tagset design for the manual annotation of the Tigrinya corpus contained 73 tags (tagset-1) with major category or level-1, subcategories or level-2, and clitics or level-3 (preposition and/or conjunction) information.

Normally, the distribution of these tags in the corpus is not balanced. For example nouns constitute about 33% of the words in the corpus whereas many other tags are rarely found. In fact, seven of the tags were not assigned to words in the corpus, while 19 tags were rarely present, each appearing fewer than 10 times in the corpus. Most of these rare tags are more complex, level-3 tags, which represent words that are cliticized with a preposition, a conjunction, or both. Therefore, in order to reduce data sparseness, another set of tags was designed by omitting level-3 annotation. After this reduction, the number of the remaining tags with level-1 and level-2 information was 20 (tagset-2). In comparison with tagset-1, the corpus with the reduced tagset had a better distribution of tags, and all of the tags were found more than 100 times in the corpus. To analyze the impact of the two tagsets on performance, experiments with both tagset-1 and tagset-2 were run. In both algorithms, the results of the reduced tagset-2 slightly outperform the larger tagset-1. To see if these results attain statistical significance, the CV results of the SVM-based tagger were considered with 'all' features (table 3). The significance test shows that the difference is statistically significant (*p*- value = 0.001 < 0.05).

Table 3: Overall performance for the data with tagset-1 and tagset-2. Context = word-2, pos-2, word-1, pos-1, word, word+1, word+2, ptn = consonant-vowel pattern, pref = prefix, suf = suffix, all = context + all affixes

Features	CRF accuracy (%)		SVM accuracy (%)	
	Tagset-1	Tagset-2	Tagset-1	Tagset-2
all	90.37	90.89	89.12	89.92
Context + pref	86.5	89.04	85.56	88.38
Context + suf	82.94	84.67	82.56	84.41
Context + ptn	81.92	83.45	82.56	84.05
Context	77.16	79.25	78.18	80.03
word (baseline)	76.15	75.15	74.05	76.43

6.4 The Effect of Prefix and Suffix Features

Considering CRF tagset-1 classifier and the features 'context+suf' and 'context+pref', the overall improvement is 5.78% and 9.34% percentage points, respectively (table 3). The impact of these features is even more visible upon detailed analysis of performance with regard to unknown words in table 4.

The result is based on a held-out evaluation in which about 81.4% of the test data are known and 18.6% are unknown words. When using only 'context' features, the performance was as low as 38.38% for the SVM-based tagger and 39.21% for the CRF-based tagger. Nevertheless, augmenting context features with affixes almost doubled this result to 76.68% and 80.22% for SVM and CRF, respectively.

Table 4: The effect of morphological features on known (kno.) and unknown (unk.) word tagging based on accuracy results (%) for tagset-1

Features	CRF		SVM		Error rate (%) of unknown words (CRF)						
	kno.	unk.	kno.	unk.	N	N_V	V_REL	V_IMF	V_AUX	ADJ	ADV
all	95.42	80.22	94.18	76.68	11.72	2.50	7.34	10.68	25.00	16.44	42.86
Context + pref + suf	95.32	79.93	94.2	75.35	10.88	5.00	5.50	11.17	37.50	17.81	42.86
Context + pref + ptn	94.45	71.44	93.91	68.49	12.55	10.00	5.50	13.59	25.00	19.18	64.29
Context + suf + ptn	92.85	60.79	93.93	57.57	16.74	47.50	42.20	36.41	12.50	21.92	71.43
Context + pref	93.21	66.98	93.35	64.43	16.32	7.50	6.42	13.59	25.00	26.03	57.14
Context + suf	91.88	56.04	92.83	53.12	17.15	77.50	47.71	34.47	12.50	23.97	85.71
Context + ptn	91.73	53.81	93.06	51.56	21.34	37.50	44.04	34.95	25.00	21.92	78.57
Context	89.04	39.21	91.75	38.38	30.96	85.00	57.80	49.03	12.50	28.77	100.00

Therefore, affix information was very helpful in inferring the POS of unknown words. Analysis of the impact of each feature shows that the addition of prefix ('pref') features contributes to the highest gain, as compared to suffixes ('suf') and patterns ('ptn'). This is also true for the combined affix features. The setup that integrates both prefix and suffix features (context+pref+suf) improved the accuracy by a significant 8.49 percentage point compared to 'context+suf+ptn,' and 19.74 percentage point compared to 'context+pref+ptn.' This effect may be due to the distribution of some very frequent prefixes such as 'mI' for verbal nouns (N_V) and 'zI, Ite' for the relative verbs. The error analysis in table 4 clearly shows that this hypothesis holds. The error rate of N_V when using prefix features was only 7.5%; while using suffixes it was 77.5%. With reference to the affixation slot position in Tigrinya words (section 2), while conjunctions are either prefixed or suffixed, prepositions are only prefixed.

6.5 The Effect of Pattern Features

The patterns of verbs, nouns, and adjectives are very informative features because inflectional and derivational rules indicate grammatical features such as verbal nouns, relative verbs, and tense-aspect-mood features, as well as gender, person, and number attributes. Incorporating patterns into the proposed contextual feature set yields considerable performance gain of unknown words. Using bare 'context' and 'ptn,' (pattern) features the accuracy increased substantially from 39.21% to 53.81% for CRF (table 4). Although less noticeable, the impact of patterns is also visible in the combined feature sets. The performance gain from 'context+pref' to 'context+pref+ptn' is about 4.46 percentage point. However, there is a small gain from 'context+pref+suf' to 'all' features. This is probably due to the long range of character *n*-grams for prefixes (1 to 6) and suffixes (1 to 5). Unless a token is quite long, the pattern features likely are already encoded in the prefix and suffix features. Therefore, the information added from patterns may not actually be new. Generally, this achievement may indicate the extent to which infix patterns can be tailored to enrich the feature encoding of Tigrinya lexical features. These types of infix features are unique to Tigrinya and other Semitic languages. Therefore, it is expected that reinforcing the feature set with more representative patterns would positively enhance the generalization of the classifiers, and thereby the prediction accuracy of POS tags.

6.6 The Effect of Data Size

The available corpus of 72k tokens is relatively small. Out of the 72k words, around 19k (26%) are unique words. For a small corpus of 72k words, this is relatively a high number of word types, which in turn increases the proportion of unknown words during testing.

Furthermore, according to the lexical diversity (token-type ratio), a word gets repeated 3.85 times in the corpus. However, about 6% of the words are hapax words, which appear only once in the entire corpus. Hapax words that happen to be in the test data become part of unknown words, which would potentially degrade the accuracy of the tagger when the POS tag is not identified correctly. The maximum overall accuracy achieved from the experiments is 90.89% from the CRF-based tagger using tagset-2. However, this result is low compared to the state-of-the-art of well-known languages such as English. For example, the TnT tagger reported an accuracy of 96.7% trained on the English Penn treebank of 1200k tokens [3].

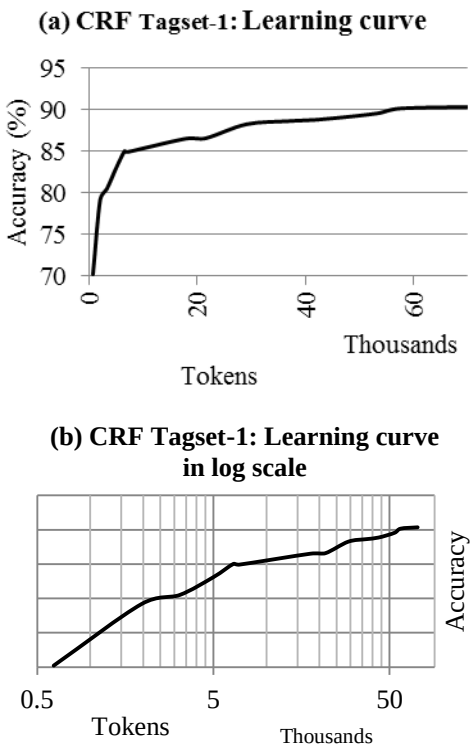

Figure 3: The effect of data size on accuracy gain

The learning curve that is depicted in figure 3 shows the accuracy of the proposed tagger growing with the availability of more data for training. The reason for this improvement is largely attributed to the reduction of out-of-vocabulary (OOV) words, or the data sparseness problem. This problem is pertinent to low-resource languages. This limitation in the training corpus was partly addressed by introducing, rich feature sets and rigorous parameter tuning to estimate the best parameters for the employed algorithms.

Although the performance of taggers across different languages and settings may not be directly comparable, data size could be one of the main reasons for the difference in performance of statistical methods. The experiments in the original implemenation of TnT [3] show the improvement of accuracy as data size is increased. To better assess the impact of data size, several experiments on successive increments of the data size at hand were run. The result from the CRF-based tagger using tagset-1 is shown in table 5.

6.7 Error Analysis

POS ambiguity in Tigrinya exists in all parts of speech. However, most of the tagging errors stem from the role of (1) relative verbs and adjectives, (2) demonstrative adjectives and pronouns, (3) adjectives and nouns, (4) imperfective verbs and auxiliary verbs, and (5) adjectives and adverbs. The right side of shows the error analysis of unknown words from selected POS tags. The overall performance of the taggers on individual tags is illustrated in the precision, recall, and F-score of the SVM-based tagger in table 6. In general, employing affix features reduces errors compared to the bare 'context' feature set. Nevertheless, the informative features for each POS tag may differ. For instance, the error rate of verbal nouns (N_V) is 7.5% using the 'context+pref' feature set, whereas it is 77.5% when using the 'context+suf' feature. This

is justified by the fact that verbal nouns are always prefixed by the morpheme ‘mI.’ The same is also true with the reduction of errors for relative verbs prefixed mostly with 'zI.' However, from the relatively low score in the precision-recall table, it’s clear that relative verbs (V_REL) were difficult to discern due to other ambiguities. One case of relative verbs, such as ‘ዝበለጸ’ /zIbeleSe/, ‘the best,’ is that they can assume the role of an adjective when modifying a noun, as in ‘ዝበለጸ ቦታ’ /zIbeleSe bota/, ‘the best place.’Prepositions and conjunctions, which form relatively few word types, are well recognized by the models. In Tigrinya, all nouns (N) may also function as adjectives (ADJ), such as when a noun modifies another noun. The estimation for noun classes has comparatively lower precision. However, the error rate figures show that joint features of 'context+pref+suf' help to lower total errors. Interestingly, while this feature set helps other tags, it does not do a better job on auxiliary verbs (table 4). The errors may be related to auxiliary verbs, such as conjugations of ነበረ /nebere/, ‘exist, live,’ ሃለወ /halewe/, ‘present,’ ጸንሐ /SenIHe/, ‘stay,’ and ተገበአ /tegebI’e/, ‘ought to,’ which are imperfective, perfective, or gerundive verbs. This type of ambiguity could partly explain the performance downgrade with categories of gerundive (V_GER) and perfective verbs (V_PRF) in table 6. In addition, auxiliary verbs form compound verbs of the pattern ‘word+V_AUX,’ which conjugate a bit differently from the four subcategories of verbs. Some examples are /hafI bele/, ‘[he] stood up,’ and /tImI belI/, ‘keep quiet.’ These constructs, annotated as ‘V,’ were predicted with considerably low precision and recall. This is due to tagging confusion with auxiliary verbs, adverbs (ADV), or the other types of verbs. In general, affixes or patterns have significant performance improvements on the POSs that undergo inflection, but limited impact on other POSs that do not get inflected.

Table 6: Precision, recall, and F1-score for the SVM – based tagger, tagset-2

POS	Precision	Recall	F1-score	Support
ADJ	0.85	0.83	0.84	870
ADV	0.79	0.79	0.79	233
CON	0.98	0.96	0.97	506
FW	0.60	0.71	0.65	17
INT	0.90	0.75	0.82	12
N	0.93	0.97	0.95	1860
NUM	0.98	0.95	0.97	120
N_PRP	0.94	0.86	0.90	244
N_V	0.97	0.99	0.98	203
PRE	0.95	0.95	0.95	447
PRO	0.83	0.85	0.84	223
PUN	1.00	1.00	1.00	778
UNC	1.00	0.67	0.80	9
V	0.69	0.65	0.67	37
V_AUX	0.95	0.96	0.96	348
V_GER	0.87	0.94	0.90	257
V_IMF	0.94	0.93	0.93	461
V_IMV	0.87	0.49	0.62	41
V_PRF	0.82	0.74	0.78	160
V_REL	0.84	0.85	0.85	382
Average	0.92	0.92	0.92	7208

7. CONCLUSION AND FUTURE WORK

This POS-tagging research based on the newly created Nagaoka Tigrinya Corpus lays out a promising foundation for initiating NLP research for Tigrinya, a low-resource and morphologically complex Semitic language. A news text corpus of 72k tokens was manually annotated for POS. Subsequently, the morphological information embedded in Tigrinya words was utilized to improve the accuracy of the tagger, particularly when annotating out-of-vocabulary words. This was performed by encoding features of prefixes, suffixes, and morphological patterns that augment the contextual word and POS features. CRF and SVM algorithms were employed, tuned through rigorous parameter optimization methods. The highest accuracy of the CRF-based tagger with 20 tags was 90.89%. This CRF tagger outperformed its SVM counter part by a statistically significant 1 percentage point ($p = 0.002 < 0.05$). The approach followed word-level tagging, as opposed to morpheme-level tagging, because prior to POS tagging, morpheme-level tagging would require morphological segmentation of words, which is currently not available for Tigrinya.

Future work will concentrate on the following primary objectives. First, there is a goal of improving the quality of the corpus and enriching its size by incorporating extra genres, in addition to news, to suit the ensuing enhancement. This will also include revising the tagset design and rectifying tagging errors and inconsistencies. It would be interesting to use recent advances of labeling strategies such as active learning to enlarge the corpus at less cost. The second objective is to replace the current word-level tagging by a morpheme-level approach, because the latter reduces the POS tags into simple POS categories by masking several layers of morphological inflections. Future work will investigate a segmentation-based tagger that is aware of the ambiguity in morpheme segmentation. In a similar path, [18] researched the implementation of a morphological analyzer and generator for Tigrinya verbs using finite state transducers. This system may be combined with the system proposed in this article for improving POS disambiguation performance. Finally, there is room to investigate the use of semi-superivised approaches of POS tagging, employing the unannotated version of the corpus to improve accuracy.

In the future, the tagger will be an essential prerequisite for further NLP studies such as base phrase chunking, syntactic parsing, machine translation, and text summarization.

8. ACKNOWLEDGMENTS

Special thanks go to Prof. Yoshiki Mikami for all his support in this research. We also thank the anonymous reviewers for their constructive comments.

9. REFERENCES

[1] Ali, B. B., and Jarray, F. 2013. Genetic approach for Arabic part of speech tagging. In International Journal on Natural Language Computing, IJNLC Vol. 2, No. 3. AIRCC.

[2] Bar-haim, R., Sima'an, K., and Winter, Y. 2008. Part-of-speech Tagging of Modern Hebrew Text. Natural Language Engineering, 14(2):223--251.

[3] Brants, T. 2000. TnT: a statistical part-of-speech tagger. Proceedings of the sixth conference on Applied Natural

Language Processing, pages 224--231.

[4] Cortes, C., and Vapnik, V. 1995. Support-vector networks. Machine Learning, pages 273--297.

[5] David, G., and Walker, K. 2001. Arabic newswire part 1 -Linguistic Data Consertium. https://catalog.ldc.upenn.edu/LDC2001T55. Accessed: 2014-10-16.

[6] Demeke, G. A., and Getachew, M. 2006. Manual annotation of Amharic news items with part-of-speech tags and its challenges. Addis Ababa. ELRC Working Papers, 2:1–17.

[7] Diab, M., Hacioglu, K., and Jurafsky, D. 2004. Automatic tagging of Arabic text: from raw text to base phrase chunks. In Human Language Technologies; 5th Meeting of the North American Chapter of the Association of Computational Linguistics, pages 149--152. Association for Computational Linguistics.

[8] Gambäck, B., Olsson, F., Argaw A. A., and Asker, L. 2009. Methods for Amharic Part-of-Speech Tagging. Proceedings of the First Workshop on Language Technologies for African Languages, (March):104--111.

[9] Gasser, M. 2012. HornMorpho 2.5 user's guide. Indiana University, Indiana.

[10] Gebre, B. G. 2010. Part of speech tagging for Amharic. Master's thesis, University of Wolverhampton.

[11] Adi, G. 2000. Tigrinya Grammar. Admas Forlag, Stockholm, 2 edition.

[12] Lafferty, J. D., McCallum, A., and Pereira, F. C. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01), pages 282--289, San Francisco, CA, USA.

[13] Khoja, S. 2001. APT: Arabic Part-of-speech tagger. In Proceedings of the Student Workshop at NAACL-2001, pages 20--25.

[14] Sebhatu, G. K. 1997. The basic principles of Tigrinian Language. ForfattaresBokmaskin, Stockholm.

[15] Maamouri, M. 2003. Arabic Treebank v.1. Linguistic Data Consortium, https://catalog.ldc.upenn.edu/LDC2001T55. Accessed: 2014-10-16.

[16] Marsi, E., Van Den Bosch, A., and Soudi, A. 2005. Memory-based Morphological Analysis and Part-of-speech Tagging of Arabic. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages 1--8, Ann Arbor. Association for Computational Linguistics.

[17] Mason, J. 1996. Tigrinya grammar. The Red Sea Press, Inc., New Jersey, 1996.

[18] Gasser, M. 2009. Semitic morphological analysis and generation using finite state transducers with feature structures. In Proceedings of the 12th conference of the European Chapter of the ACL, page 309–317. ACL.

[19] Mohammed, E. and Kübler, S. 2010. Is Arabic part of speech tagging feasible without word segmentation? In Human Language Technologies; 5th Meeting of the North American Chapter of the Association of Computational Linguistics, pages 705--708. Association for Computational Linguistics.

[20] Habash, N., and Rambow, O. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one full swoop. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL'05, pages 573--580, Stroudsburg, PA,USA. Association for Computational Linguistics.

[21] Omer, O., and Mikami, Y. 2012. Stemming Tigrinya Words for Information Retrieval. In Proceedings of COLING 2012: Demonstration Papers, pages 345--352, Mumbai.

[22] Amanuel, S. 1998. A Comprehensive Tigrinya Grammar. The Red Sea Press, Inc., Lawrenceville NJ.

[23] Savova, V., and Peshkin, L. 2003. Part-of-speech tagging with minimal lexicalization. In Proceedings of CoRR, 2003.

[24] Streiter, O., Scannell, K., and Stuflesser, M. 2007. Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for developers. Springer Science+Business Media.

[25] Daniel, T. R. 2005. Modern grammar of Tigrinya language. Mega Publishing and Distribution PLC, Addis Ababa.

[26] Tseng, H., Jurafsky, D., and Manning, C. 2005. Morphological features help POS tagging of unknown words across language varieties. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, 2005. Asian Federation of Natural Language Processing.

[27] Tedla, Y. K., Yamamoto, K. and Marasinghe, A. 2016. Nagaoka Tigrinya Corpus: Design and Development of Part-of-speech Tagged Corpus. In Language Processing Society 22nd Annual Meeting Papers Collection, Tohoku, Japan,. The Association for Natural Language Processing.